All About Data

Dr. Abhijnan Chakraborty
Department of Computer Science & Engg.,
Indian Institute of Technology Kharagpur
https://cse.iitkgp.ac.in/~abhijnan

What is Data?

- □ Collection of **data objects** and their **attributes**
- □ An **attribute** is a property of an object
  - – Examples: eye color of a person, temperature, etc.
  - – Attribute is also known as variable, field, characteristic, dimension, or feature
- □ A collection of attributes describe an **object**
  - – Object is also known as record, point, case, sample, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Loan?**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters

  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, pin codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius.
  - Ratio
    - Examples: length, counts, elapsed time

Properties of Attribute Values

- □ The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: $= \ \neq$
  - Order: $< \ >$
  - Differences are meaningful : $+ \ -$
  - Ratios are meaningful $* \ /$

- □ Nominal attribute: distinctness
- □ Ordinal attribute: distinctness & order
- □ Interval attribute: distinctness, order & meaningful differences
- □ Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° Celsius?

- Consider measuring the height above average
  - If Bill’s height is three inches above average and Bob’s height is six inches above average, then would we say that Bob is twice as tall as Bill?

	Attribute Type	**Description**	**Examples**	**Operations**
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ2 test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Discrete and Continuous Attributes

□ **Discrete Attribute**

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Binary attributes are a special case of discrete attributes

□ **Continuous Attribute**

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- □ Only presence (a non-zero attribute value) is regarded as important
  - ➢ Words present in documents
  - ➢ Items present in customer transactions

- □ If we met a friend in a grocery store, would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."

Types of Datasets

- □ Record
  - – Data Matrix
  - – Document Data
  - – Transaction Data
- □ Graph
  - – World Wide Web
  - – Molecular Structures
- □ Ordered
  - – Spatial Data
  - – Temporal Data
  - – Sequential Data
  - – Genetic Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Loan?
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

- If data objects have the same set of numeric attributes, we can think of them as points in a multi-dimensional space. Each dimension of this space represents one of the attributes.

- Such a data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	**Projection of y load**	**Distance**	**Load**	**Thickness**
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- □ Each document becomes a ‘term’ vector
  - – Each term is a component (attribute) of the vector
  - – The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

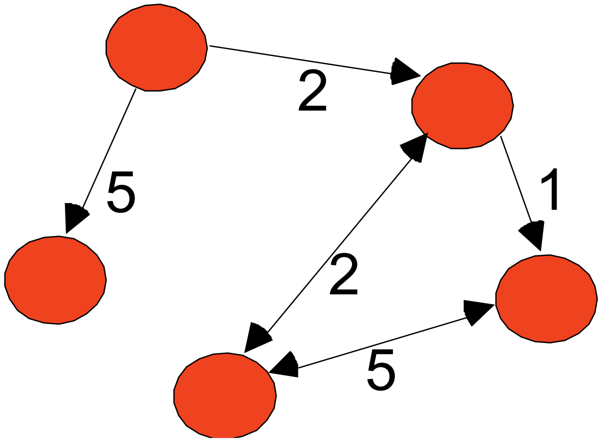
Transaction Data

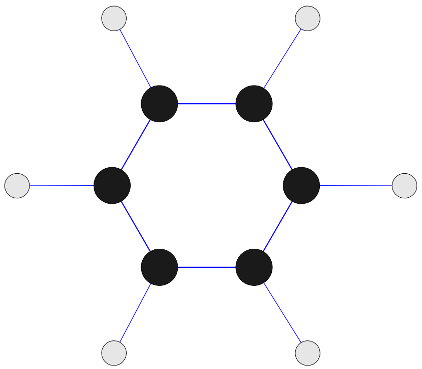
- A special type of data, where each transaction involves a set of items.
- For example, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items.
- Can represent transaction data as record data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

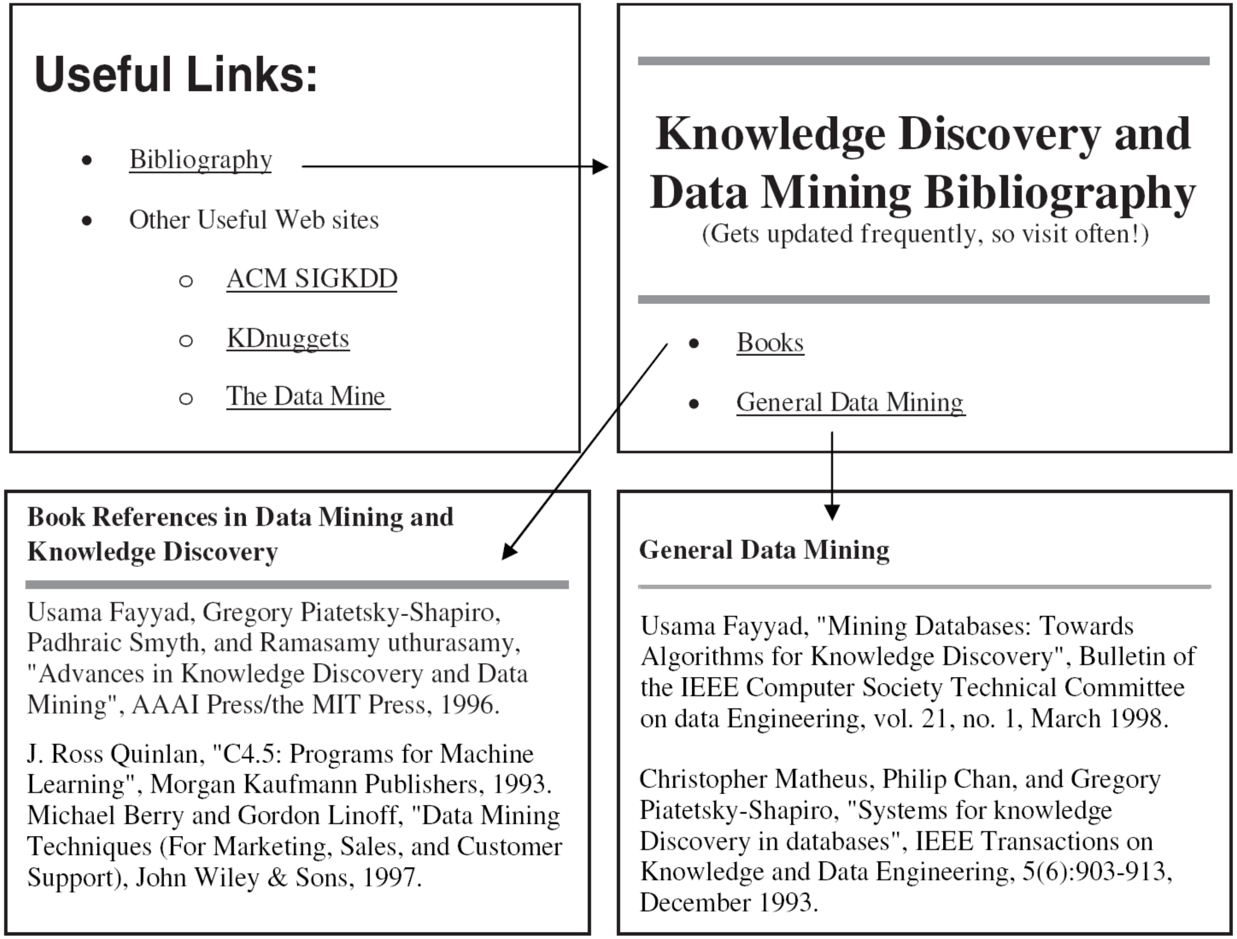
Graph Data

- □ Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6



Useful Links:

- Bibliography
- Other Useful Web sites
  - ○ ACM SIGKDD
  - ○ KDnuggets
  - ○ The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

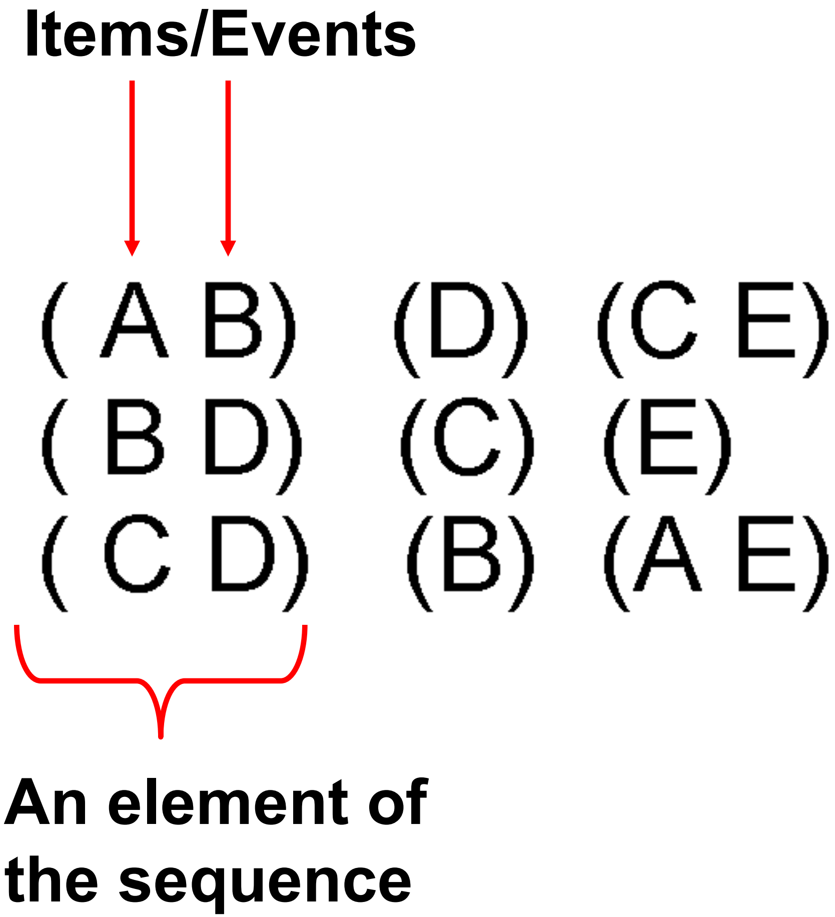
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- □ Sequences of transactions



Items/Events

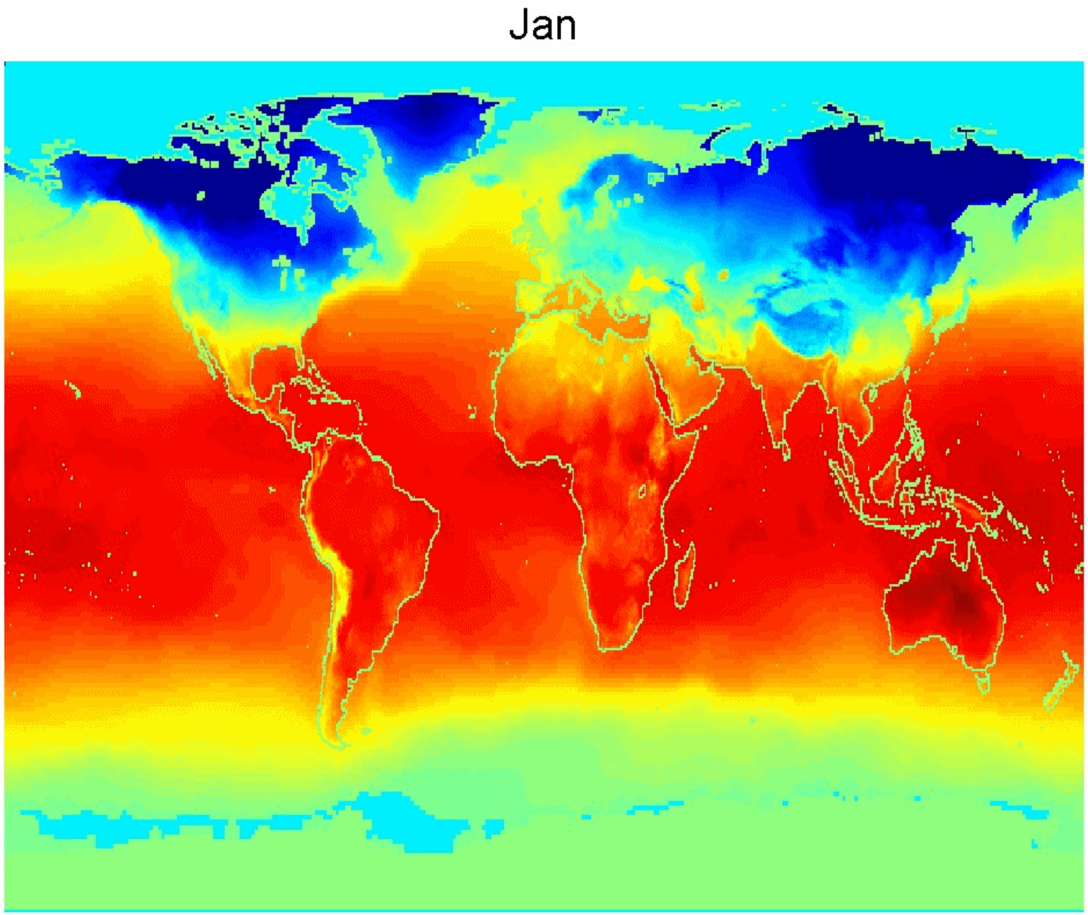
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of
the sequence

Ordered Data

- □ Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

- □ Poor data quality negatively affects many data processing efforts

- □ If a classification model for detecting people who are loan risks is built using poor data
  - – Some credit-worthy candidates are denied loans
  - – More loans are given to individuals that default

Data Quality Issues

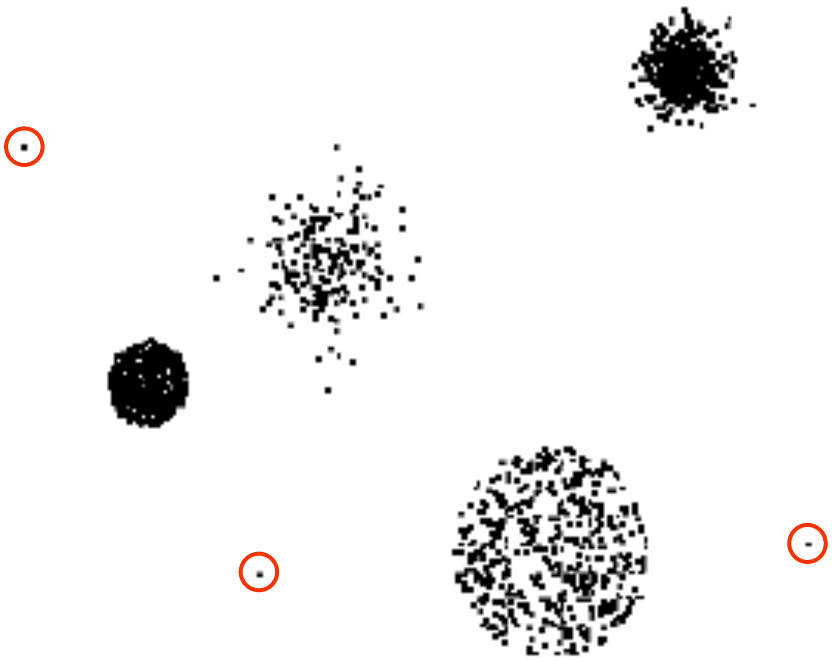
- Noise and outliers
- Wrong data
- Fake data
- Missing values
- Duplicate data

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone

Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



Missing Values

- □ Reasons for missing values
  - – Information is not collected
    (e.g., people decline to give their age and weight)
  - – Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- □ Handling missing values
  - – Eliminate data objects or variables. [Deletion]
  - – Estimate missing values [Imputation]
    - ➢ Example: time series of temperature
    - ➢ Example: census results
  - – Ignore the missing value during analysis

Handling Missing Data (Specify Your Assumptions)

- 5% of student records at a university have no “marital status” (single, married, ...)
  - Drop records? Impute value, how?
- 5% of smokers in a study of the effects of tobacco on health had no year of birth
  - Drop records? Impute value, how?
- 5% of records of sales of a company have pin code but no state
  - Drop records? Impute value, how?
- Temperature sensor at weather station was failing at random intervals for one day, total downtime 6 hours, max continuous downtime 15 minutes
  - Drop that day? Impute values, how?
- Same sensor failed for one night, downtime 6 hours continuous
  - Drop that day? Impute values, how?

Possible Answers

- 5% of student records at a university have no "marital status"
  - Undergrads? Impute as "single" unless there is a "spouse" field
- 5% of smokers in a study of the effects of tobacco on health had no year of birth
  - Drop, but check if there is something systematic in distribution of other values for them
- 5% of records of sales of a company have pin code but no state
  - Get a table for zip code to state; complete the missing data
- Temperature sensor at weather station was failing at random intervals for one day, total downtime 6 hours, max continuous downtime 15 minutes
  - Impute by interpolating
- Same sensor failed for one night, downtime 6 hours continuous
  - Drop that day; interpolation may be inaccurate

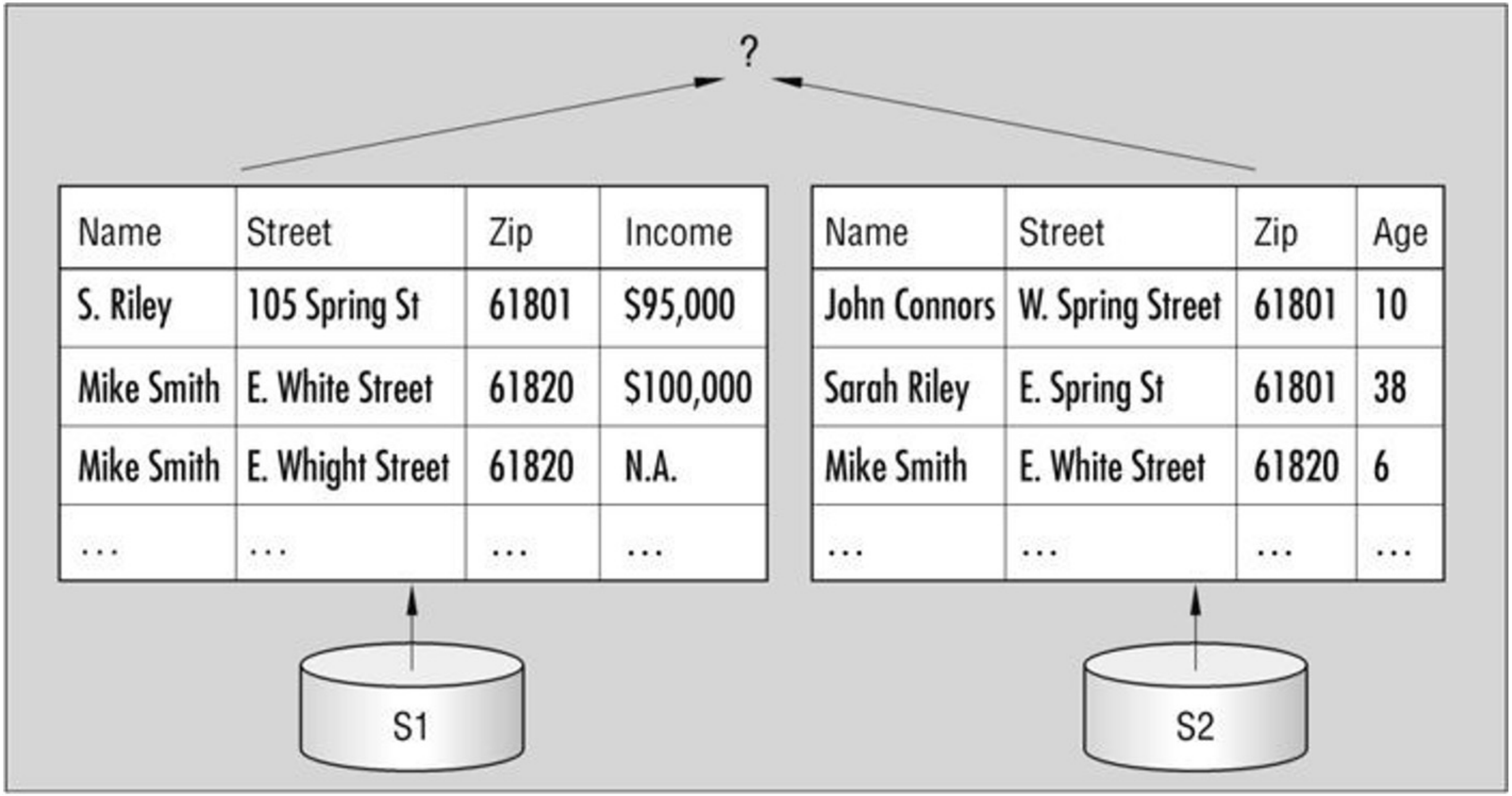
Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

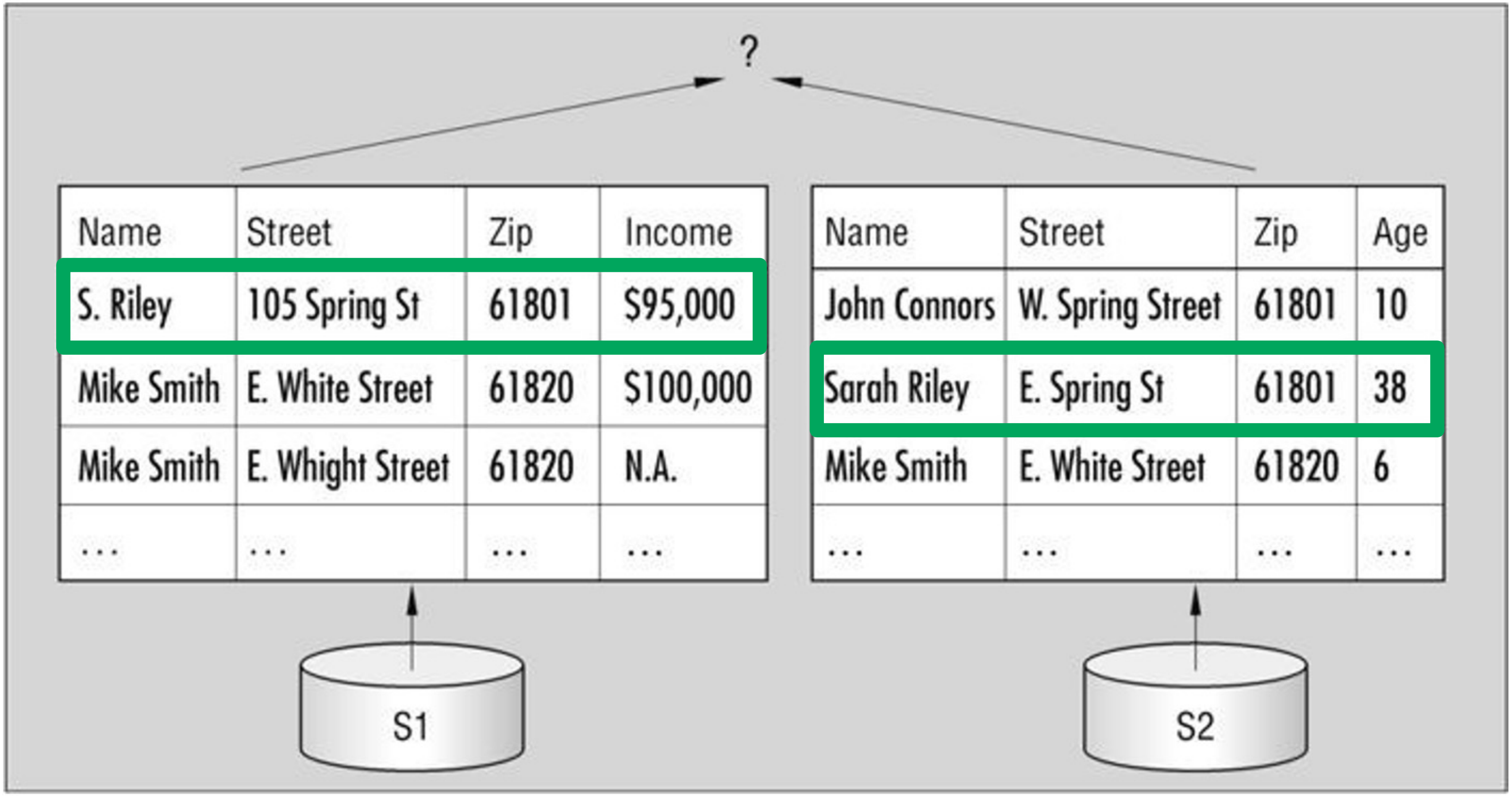
Data integration is not easy



Name	Street	Zip	Income
S. Riley	105 Spring St	61801	$95,000
Mike Smith	E. White Street	61820	$100,000
Mike Smith	E. Whight Street	61820	N.A.
...	...	...	...

Name	Street	Zip	Age
John Connors	W. Spring Street	61801	10
Sarah Riley	E. Spring St	61801	38
Mike Smith	E. White Street	61820	6
...	...	...	...

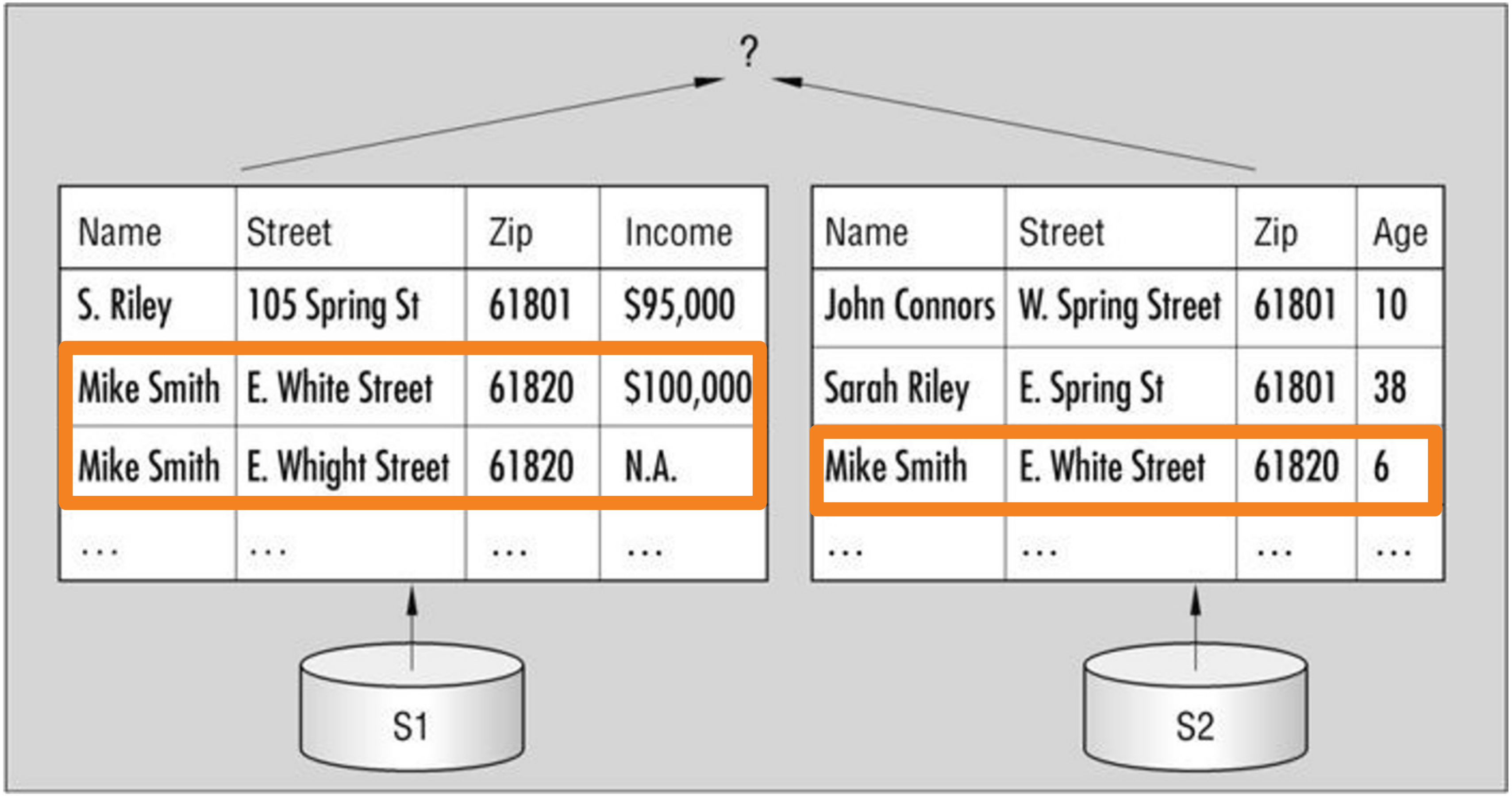
Data integration is not easy



Name	Street	Zip	Income
S. Riley	105 Spring St	61801	$95,000
Mike Smith	E. White Street	61820	$100,000
Mike Smith	E. Whight Street	61820	N.A.
...	...	...	...

Name	Street	Zip	Age
John Connors	W. Spring Street	61801	10
Sarah Riley	E. Spring St	61801	38
Mike Smith	E. White Street	61820	6
...	...	...	...

Data integration is not easy



Name	Street	Zip	Income
S. Riley	105 Spring St	61801	$95,000
Mike Smith	E. White Street	61820	$100,000
Mike Smith	E. Whight Street	61820	N.A.
...	...	...	...

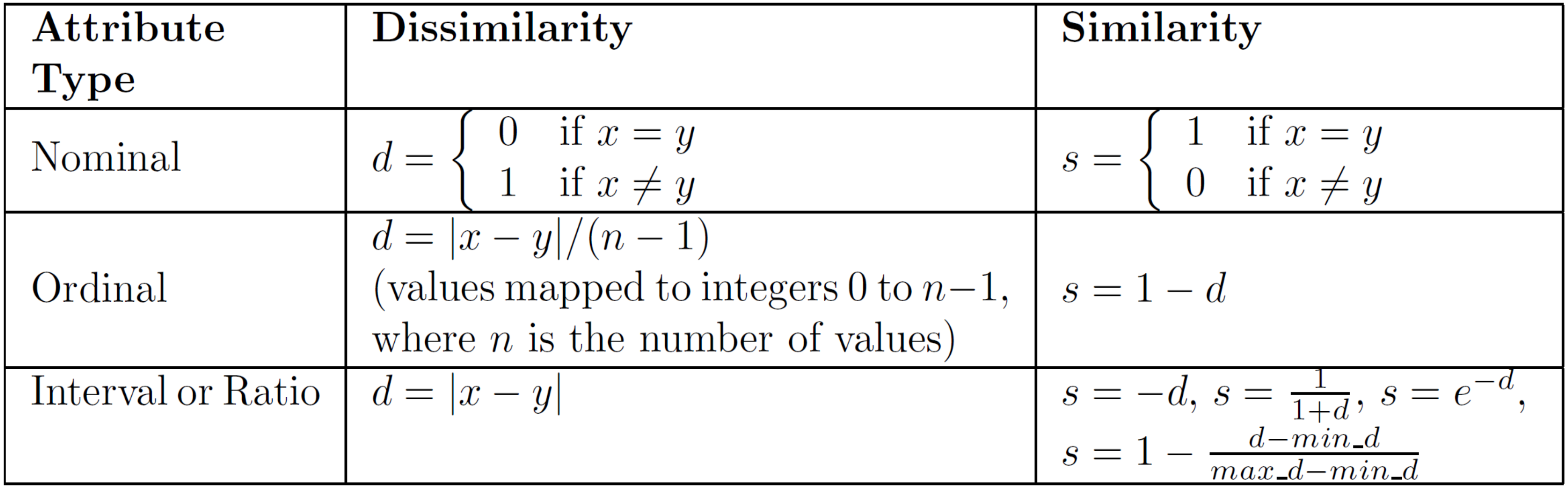
Name	Street	Zip	Age
John Connors	W. Spring Street	61801	10
Sarah Riley	E. Spring St	61801	38
Mike Smith	E. White Street	61820	6
...	...	...	...

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Higher when objects are more alike.
  - Often falls in the range [0,1]

- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

Similarity/Dissimilarity for Simple Attributes



Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \|x - y\|/(n-1)$ (values mapped to integers 0 to $n{-}1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \|x - y\|$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

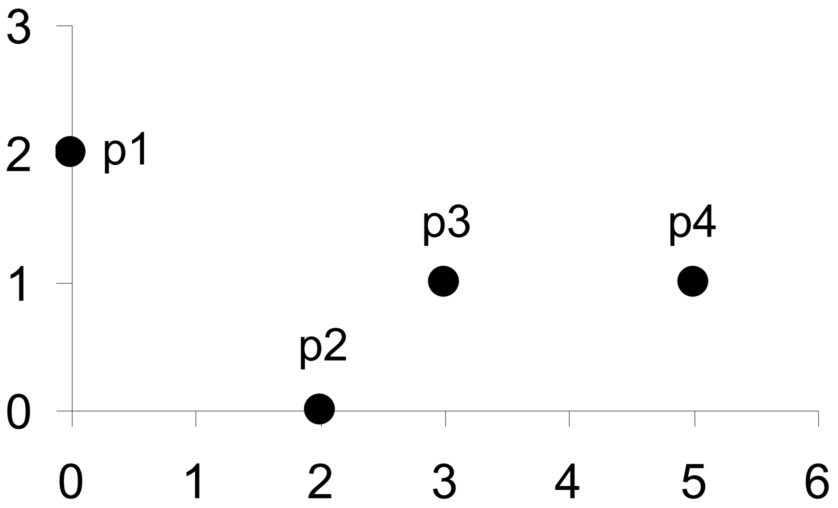
- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are the kth attributes (components) of **x** and **y**.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

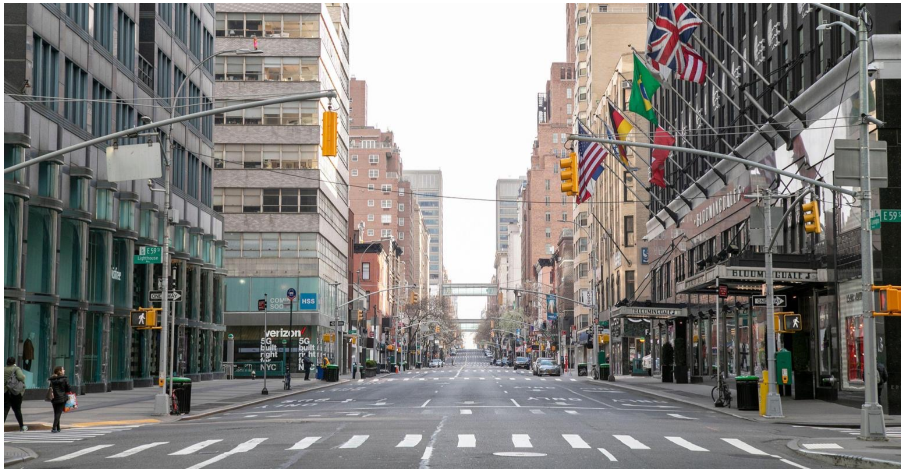
□ Minkowski Distance is a generalization of Euclidean Distance

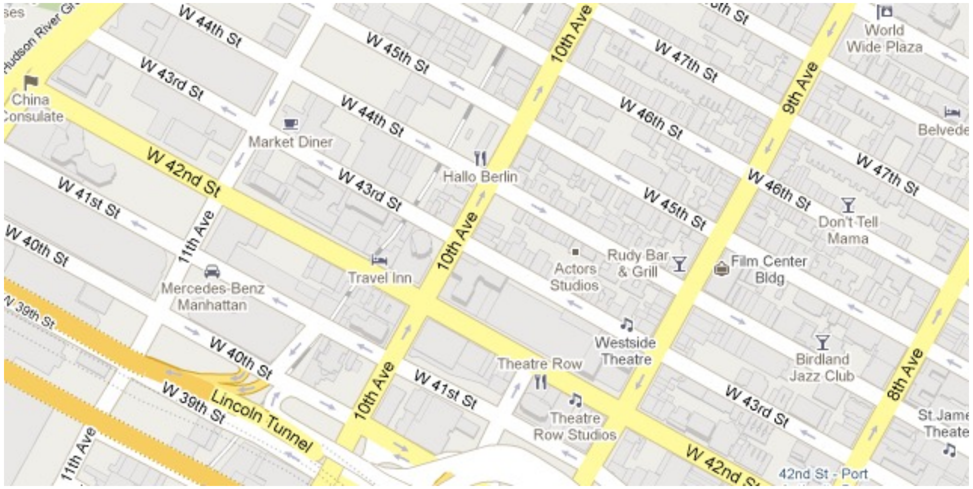
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the kth attributes (components) or data objects x and y.

Minkowski Distance: Examples

- □ $r = 1$. City block (Manhattan, L_1 norm) distance.
  - – A common example for binary vectors is the Hamming Distance, which measures how many bits are different between two vectors





Minkowski Distance: Examples

- □ $r = 1$. City block (Manhattan, L_1 norm) distance.
  - – A common example for binary vectors is the Hamming Distance, which measures how many bits are different between two vectors

- □ $r = 2$. Euclidean distance

- □ $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - – This is the maximum difference between any component of the vectors

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

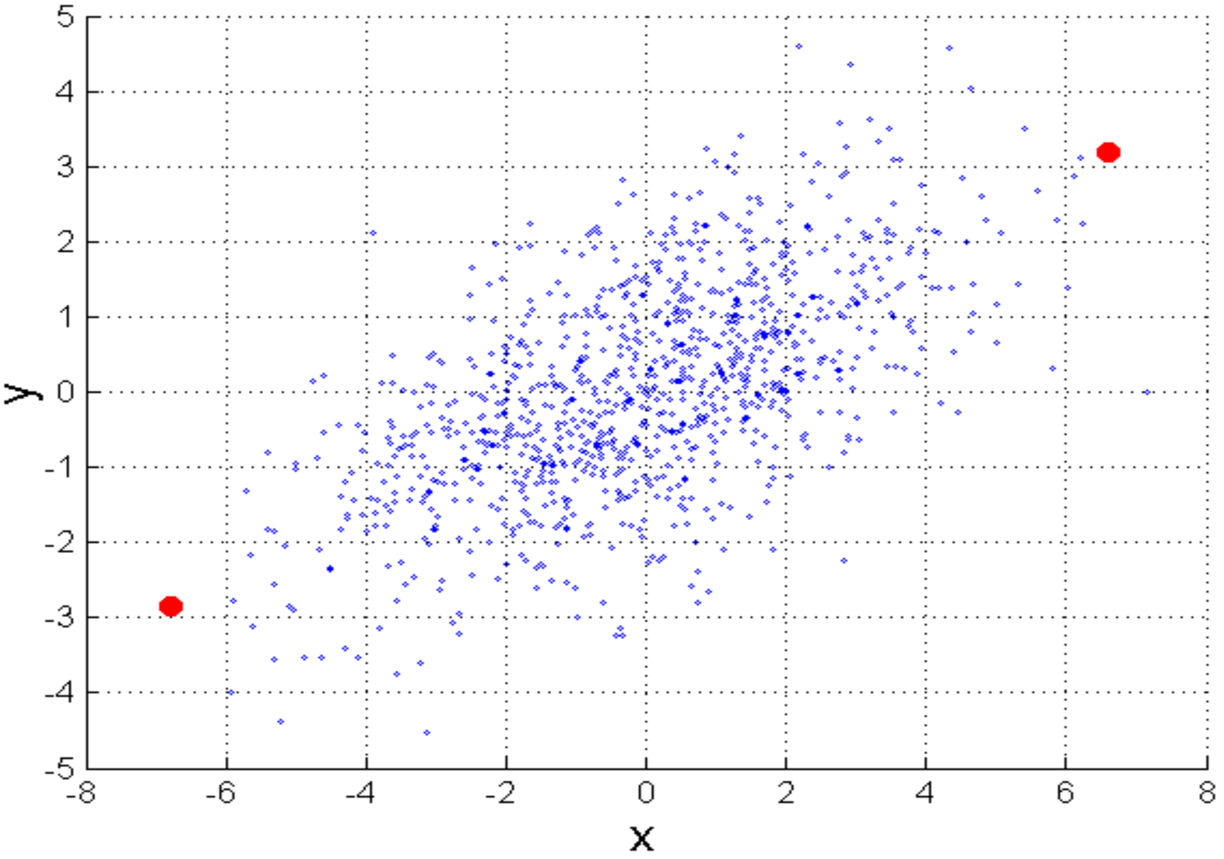
L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Mahalanobis Distance

$$\mathbf{mahalanobis}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$



Σ is the covariance matrix

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well-known properties.

  1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$.
  2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all **x** and **y**. (Symmetry)
  3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points **x**, **y**, and **z**. (Triangle Inequality)

  where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), **x** and **y**.

- A distance that satisfies these properties is called a metric

Similarity Between Binary Vectors

- □ Common situation is that objects, **x** and **y**, have only binary attributes

- □ Compute similarities using the following quantities

  f_{01} = # attributes where **x** was 0 and **y** was 1

  f_{10} = # attributes where **x** was 1 and **y** was 0

  f_{00} = # attributes where **x** was 0 and **y** was 0

  f_{11} = # attributes where **x** was 1 and **y** was 1

- □ Simple Matching Coefficient = number of matches / number of attributes = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  Jaccard Coefficient = number of 11 matches / number of non-zero attributes = $(f_{11}) / (f_{01} + f_{10} + f_{11})$

SMC vs. Jaccard: Example

$\mathbf{x} =$ 1 0 0 0 0 0 0 0 0 0

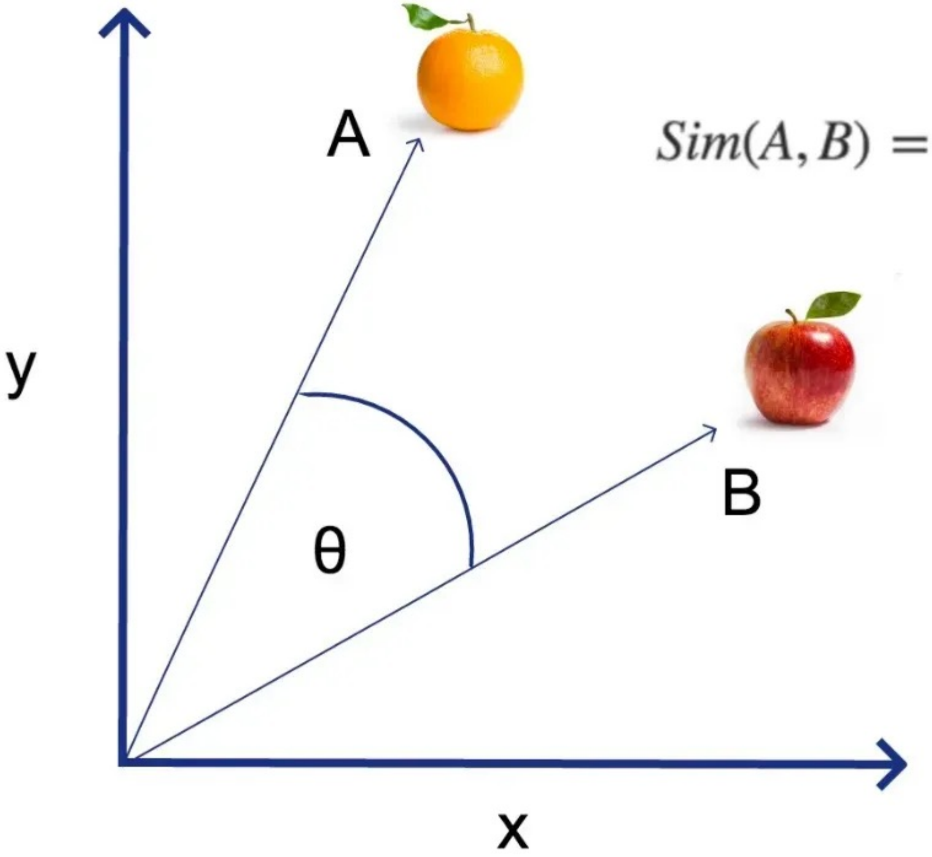
$\mathbf{y} =$ 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2 \mid f_{10} = 1 \mid f_{00} = 7 \mid f_{11} = 0$

$$\begin{aligned} \text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7 \end{aligned}$$

$\text{JC} = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine Similarity



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Cosine Similarity

□ If **d**$_1$ and **d**$_2$ are two document vectors, then

$$cos(d_1, d_2) = \frac{< d_1, d_2 >}{||d_1|| \cdot ||d_2||}$$

where <**d**$_1$,**d**$_2$> indicates dot product of vectors **d**$_1$ and **d**$_2$

||d|| is the length of vector **d**.

□ Example:

d$_1$ **= 3 2 0 5 0 0 0 2 0 0**

d$_2$ **= 1 0 0 0 0 0 0 1 0 2**

<d$_1$**, d**$_2$**>** = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5

|| **d**$_1$ || = √(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0) = 6.481

|| **d**$_2$ || = √(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) = 2.449

cos(**d**$_1$, **d**$_2$) = 0.3150

Correlation between Objects

- ❑ The correlation between two variables X and Y, $\text{Corr}_{(X,Y)}$ can be defined as:

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:
- Cov(X, Y) is the covariance between X and Y
- σ_X, σ_Y are standard deviation of X and Y

- ❑ Cov(X, Y) is defined as

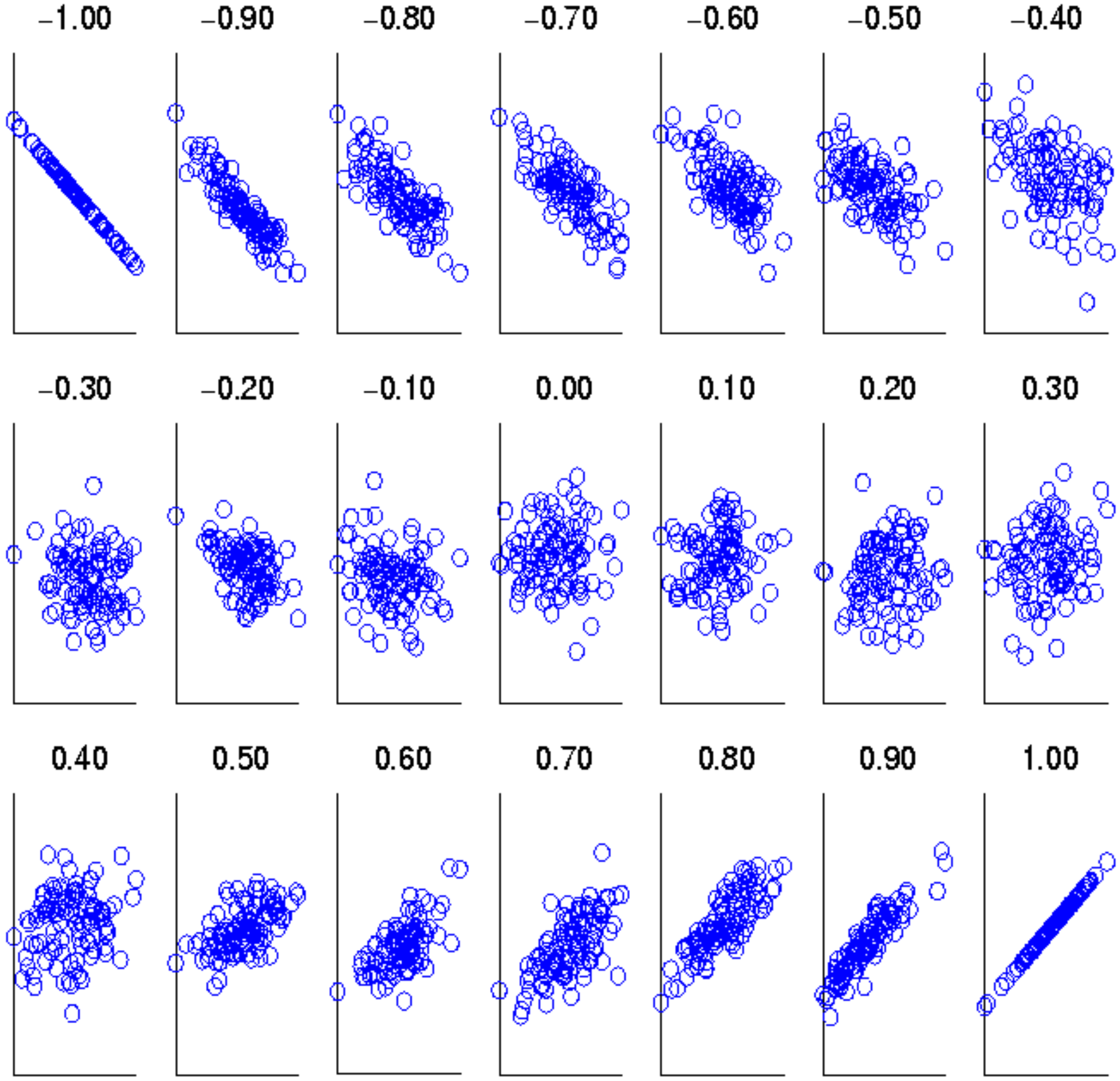
$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$\bar{X}, \bar{Y}$ are the means of X and Y respectively

- ❑ Standard deviation is defined as

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Visually Evaluating Correlation

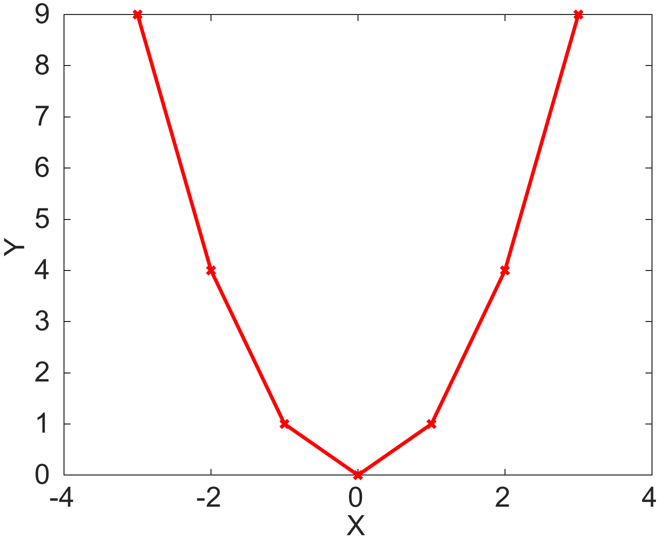


Scatter plots showing the correlation value ranging from –1 to 1

Drawback of Correlation

- □ x = (-3, -2, -1, 0, 1, 2, 3)
- □ y = (9, 4, 1, 0, 1, 4, 9)

$y_i = x_i^2$



- □ mean(x) = 0, mean(y) = 4
- □ std(x) = 2.16, std(y) = 3.74

- □ corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)
  = 0

Only linear relationship can be captured

Correlation vs Cosine vs Euclidean Distance

- Compare the three proximity measures according to their behavior under variable transformation
  - scaling: multiplication by a value
  - translation: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

Correlation vs Cosine vs Euclidean Distance

- Consider the example
  - **x** = (1, 2, 4, 3, 0, 0, 0), **y** = (1, 2, 3, 4, 0, 0, 0)
  - $\mathbf{y_s} = \mathbf{y} * \mathbf{2}$ (scaled version of y)
  - $\mathbf{y_t} = \mathbf{y} + \mathbf{5}$ (translated version)

Measure	(x, y)	(x, y_s)	(x, y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Sampling

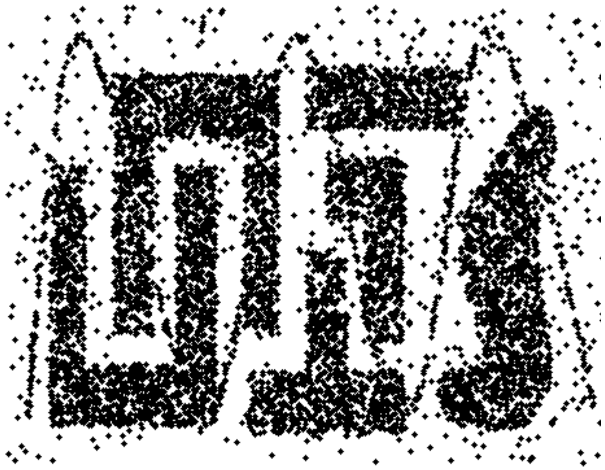
- □ Sampling is the main technique employed for data reduction.
- □ Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.
- □ Sampling is useful because processing the entire set of data of interest is too expensive or time consuming.

Sampling …

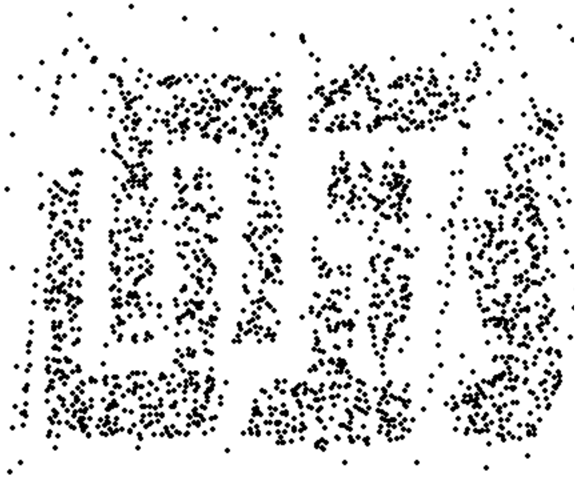
□ The key principle for effective sampling is the following:

- Using a sample will work almost as well as using the entire data set, if the sample is representative
- A sample is representative if it has approximately the same properties (of interest) as the original set of data

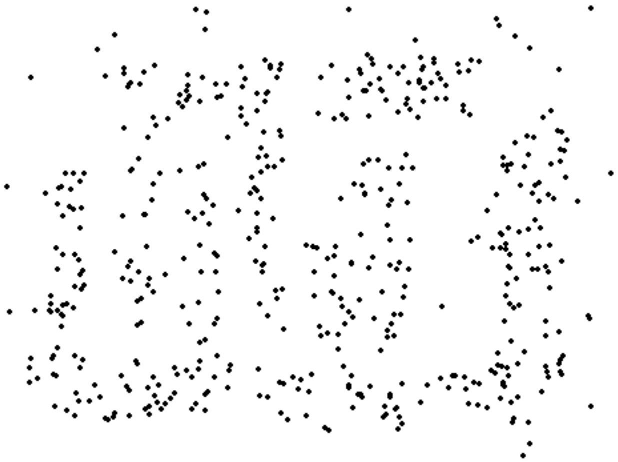
Sample Size



8000 points



2000 Points



500 Points

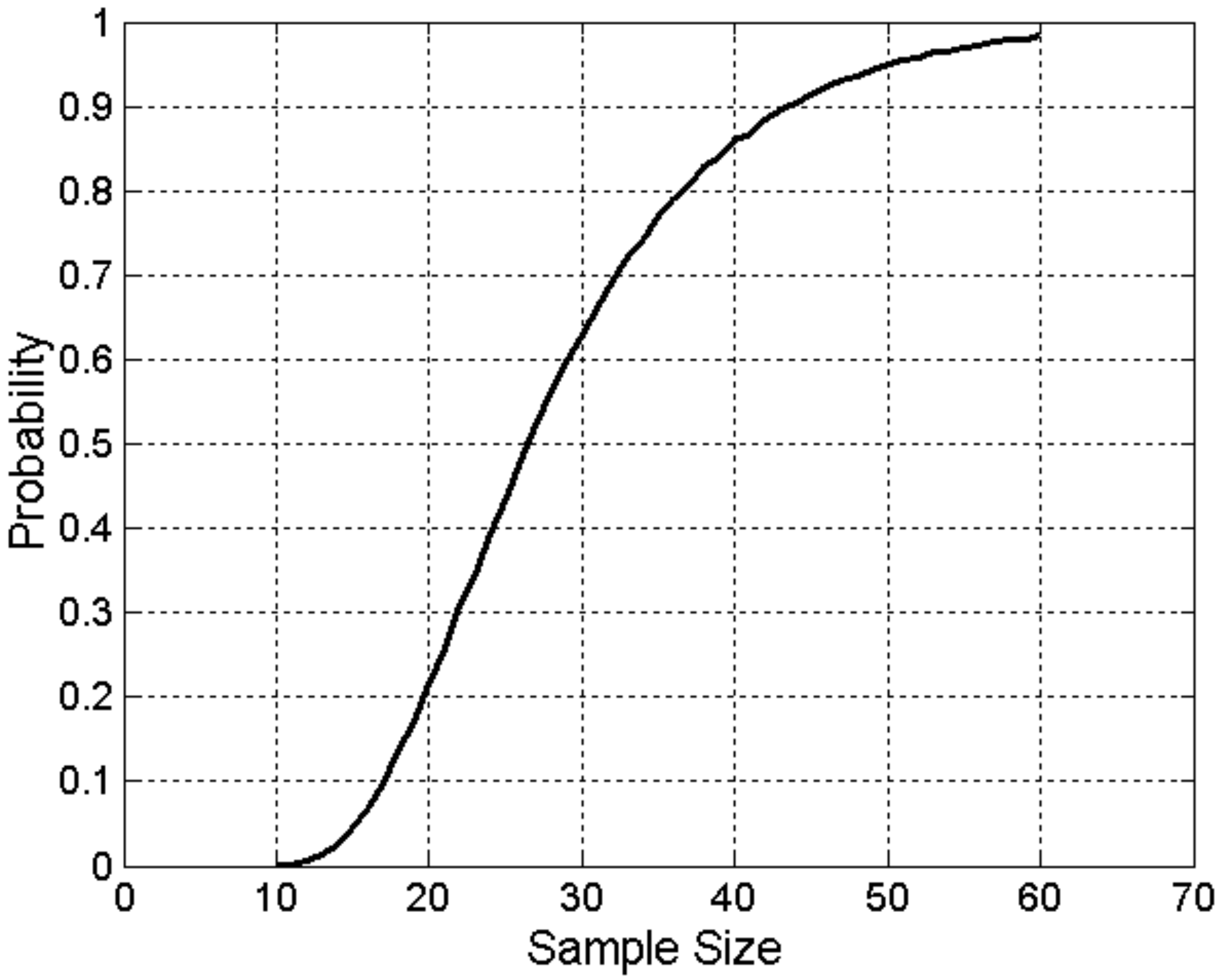
Types of Sampling

- ❑ Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - ➢ As each item is selected, it is removed from the population
  - Sampling with replacement
    - ➢ Objects are not removed from the population as they are selected for the sample
    - ➢ In sampling with replacement, the same object can be picked up more than once

- ❑ Stratified Sampling
  - Split the data into several partitions; then draw random samples from each partition

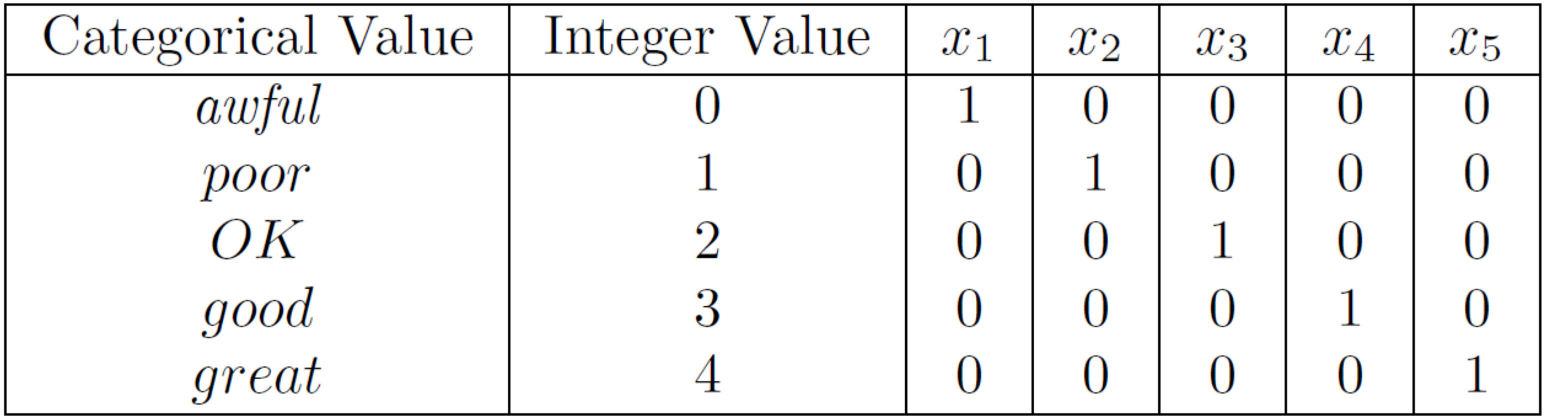
Sample Size

- □ What sample size is necessary to get at least one object from each of 10 equal-sized groups?



Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables



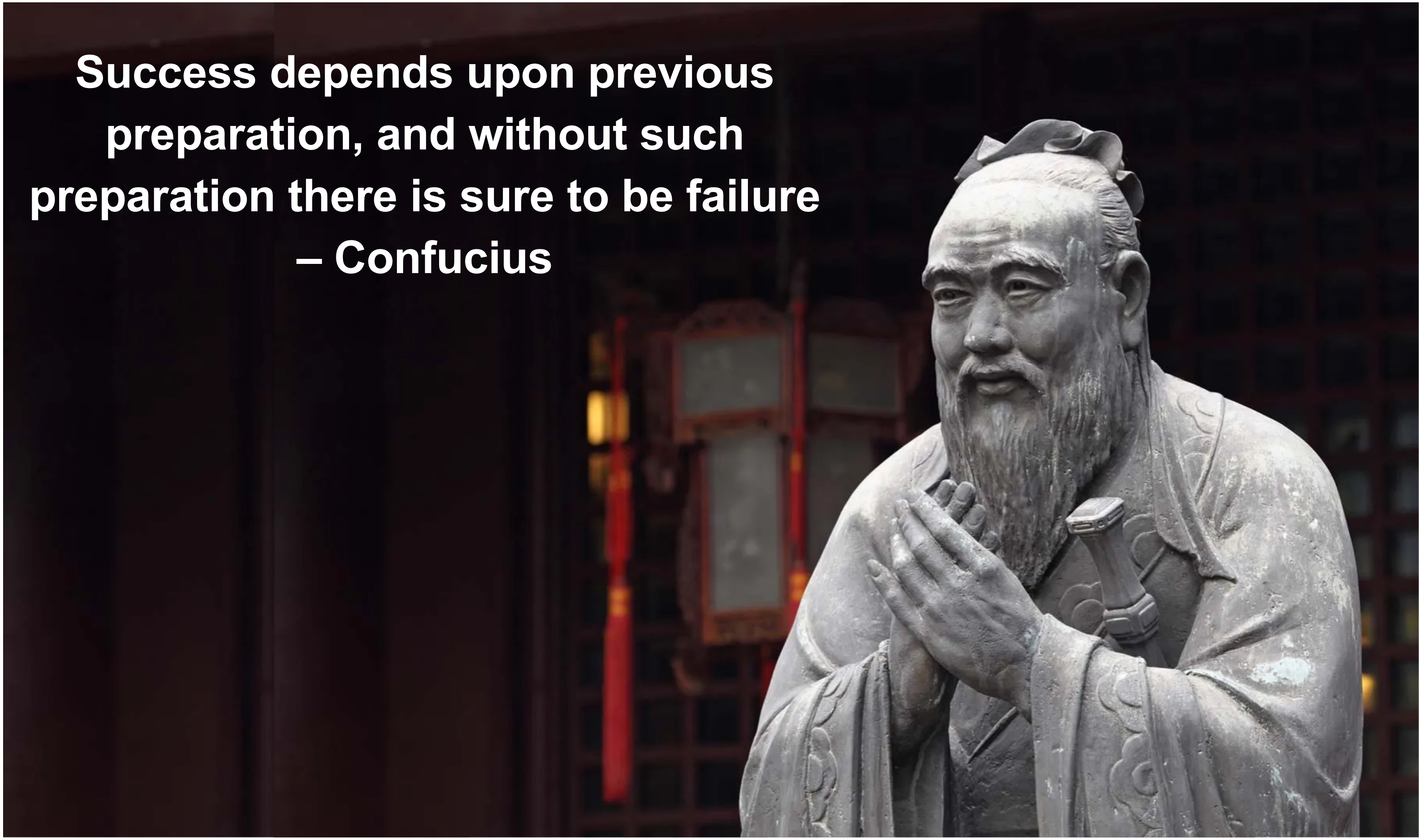
Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Attribute Transformation

- An attribute transform is a function that changes all the values of an attribute to new values, where each original value has a specific new value it corresponds to.
    - Simple functions: x^k, $\log(x)$, e^x, $|x|$

- Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Common standardization approach is subtracting off the means and dividing by the standard deviation

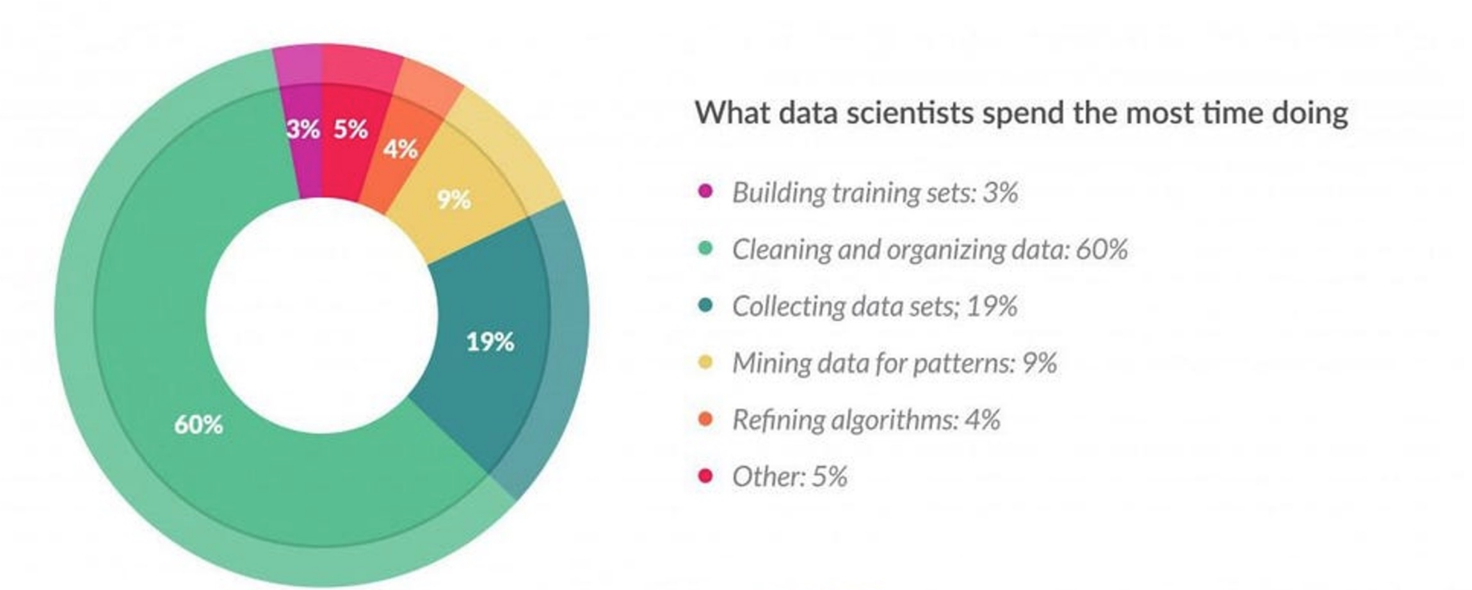
Data Preparation is Paramount!



Success depends upon previous preparation, and without such preparation there is sure to be failure

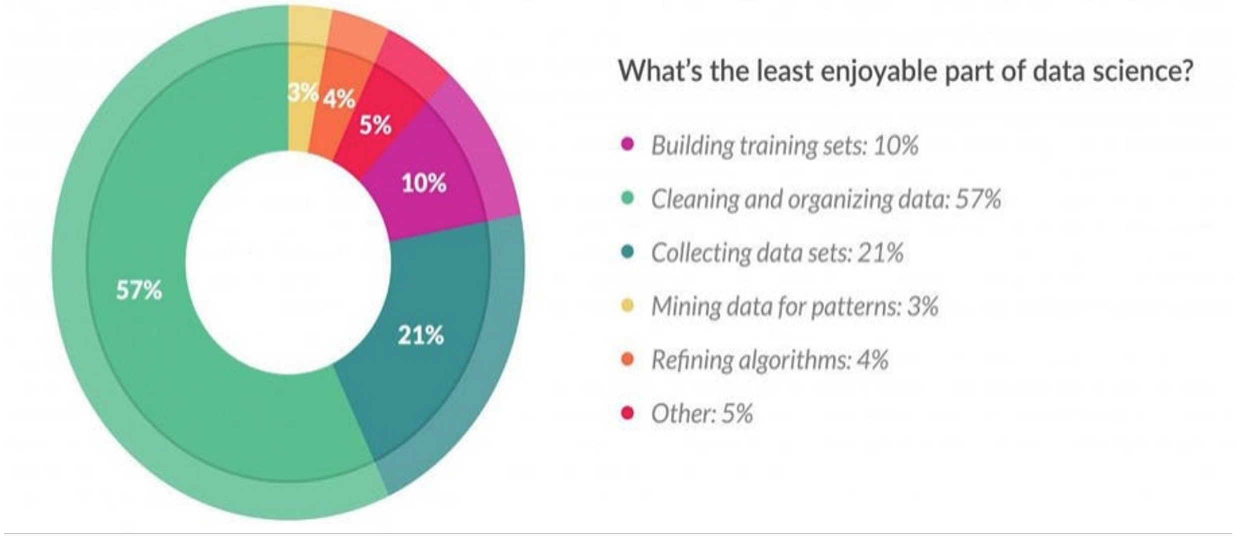
– Confucius

Preparation is time consuming



What data scientists spend the most time doing

- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*



What's the least enjoyable part of data science?

- ● *Building training sets: 10%*
- ● *Cleaning and organizing data: 57%*
- ● *Collecting data sets: 21%*
- ● *Mining data for patterns: 3%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

Forbes 2016

Thank You

Slides Courtesy
1. Introduction to Data Mining, 2nd Edition
by Tan, Steinbach, Karpatne, Kumar
2. Prof. Carlos Castillo, UPF Barcelona